A framework for learning depth from a flexible subset of dense and sparse light field views

Jinglei Shi, Xiaoran Jiang, Christine Guillemot

A framework for learning depth from a flexible subset of dense and sparse light field views

Jinglei Shi, Xiaoran Jiang, Christine Guillemot *Fellow, IEEE*

***Abstract*—In this paper, we propose a learning based depth estimation framework suitable for both densely and sparsely sampled light fields. The proposed framework consists of three processing steps: initial depth estimation, fusion with occlusion handling, and refinement. The estimation can be performed from a flexible subset of input views. The fusion of initial disparity estimates, relying on two warping error measures, allows us to have an accurate estimation in occluded regions and along the contours. In contrast with methods relying on the computation of cost volumes, the proposed approach does not need any prior information on the disparity range. Experimental results show that the proposed method outperforms state-of-the-art light fields depth estimation methods, including prior methods based on deep neural architectures.**

***Index Terms*—light fields, depth estimation, deep neural network, occlusion handling.**

I. Introduction

LIGHT fields, by recording the radiance of light rays along different orientations yield a very rich description of the scene, enabling 3D scene geometry estimation and 3D scene reconstruction. Scene depth (or equivalently disparity) estimation methods have been recently proposed for light fields that can be broadly classified in two categories. The first category of approaches analyze specific linear structures in Epipolar Plane Images (EPI) [1] [2] for depth computation from dense light fields. Indeed, the corresponding pixels in different views of a light field form a line in EPI, whose slope is proportional to the disparity value [3]. Another category of methods adopts techniques from classical stereo reconstruction, i.e., matching corresponding pixels in all sub-aperture images (SAI) or views of the light field, essentially using robust patch-based block matching [4] [5] [6]. A cost volume is constructed in [4] to evaluate the matching cost defined as similarity between the sub-aperture images and the central image shifted at different sub-pixel locations.

EPI based methods are only suitable for densely sampled light fields. While stereo methods allow estimating larger disparities, they need to discretize the disparity space to compute cost volumes, using also some prior knowledge on the disparity range. A high discretization level improves estimation accuracy but yields a heavy computational cost, leading often to estimating depth at the central viewpoint only. The authors in [6] estimate disparity maps for every viewpoint from only a subset of light field views (in particular the four corner views). These estimated disparity maps are then propagated by warping to the target view, using a low rank completion approach to cope with holes due to occlusions. The author in [5] recently proposed an empirical Bayesian framework to estimate scene-dependent parameters for inferring scene depth. In [7], by dividing a light field into several stereo image pairs, the authors estimate corresponding disparities through a multiscale and multiwindow (MSMW) stereo matching method and then process them with an optical flow based interpolation. The final disparity is obtained by a median fusion of the initial disparities.

Besides the above disparity estimation methods, a disparity map can be also estimated with an optical flow estimator, as both disparity and optical flow measure pixel displacement between two images. Recently, the field of optical flow estimation has known significant advances thanks to the use of deep neural networks trained in a supervised or unsupervised manner. An end-to-end trainable encoder-decoder network, called *FlowNet*, is proposed in [8] for optical flow estimation. The architecture is further improved in [9] by stacking several elementary networks, each of them being similar to *FlowNet*. The resulting architecture, called *FlowNet 2.0*, significantly improves the prediction accuracy and is further refined in [10] to improve the performance in occluded regions and contours. A pyramid structure for flow estimation is proposed in [11], which generates competitive results with less parameters. The authors in [12] achieve state-of-the-art results by combining a pyramid structure, warped features correlation and cost volume. Deep learning methods have also been successfully applied to many light fields processing tasks such as view synthesis [13] [14] and super-resolution [15] [16]. Although a few architectures have been proposed for scene depth (or disparity) estimation from dense light fields based on EPI [17], [18], very few methods have been proposed so far for sparse light fields.

In this paper, we propose a supervised deep learning framework for estimating scene depth, taking at the input a flexible subset of light field views. In order to compute scene depth, the proposed approach estimates disparity maps for every viewpoint of the light field. Hence, in the rest of the paper, we will refer to disparity estimation only. The use of subsets of input views allows us, compared to stereo estimation methods, to increase the estimation accuracy, while limiting computational complexity. Initial disparity estimates are computed between aligned stereo pairs using the *FlowNet 2.0* optical flow estimation architecture that we fine-tuned to be suitable for disparity estimation in dense and sparse light fields. These initial estimates are used to warp a flexible set of anchor views onto a target viewpoint. The fusion of these initial estimates relying on a *winner-takes-all* (WTA) strategy with

This project has been supported by the EU H2020 Research and Innovation Programme under grant agreement No 694122 (ERC advanced grant CLIM).

two measures of warping errors reflecting disparity inaccuracy in occlusion-free and occlusions respectively, allows us to have an accurate disparity estimation in occluded regions and along the contours. A refinement network is then proposed to learn the disparity maps residuals at different scales. The proposed new architecture relies in part on the one considered in [19], by however extending the approach into a more general framework, which enables to perform estimation from a flexible subset of input views.

The training of the proposed neural networks based architecture requires having ground truth disparity (or depth) maps. Although a few synthetic datasets exist for dense light fields with ground truth depth maps, no such dataset exists for sparse light fields with large baselines. This lack of training data with ground truth depth maps is a crucial issue for supervised learning of neural networks for depth estimation. We therefore created two datasets, called SLFD and DLFD, respectively containing sparsely and densely sampled synthetic light fields. DLFD contains light fields having a disparity range within the interval [-4,4] between adjacent views, i.e. of the same order of magnitude as light fields captured with plenoptic cameras. SLFD contains light fields with a larger disparity range, i.e. within the interval [-20,20], which is comparable to the one of light fields captured with camera rigs. To our knowledge, SLFD is the first available dataset providing sparse light field views and their corresponding ground truth depth and disparity maps. The created datasets will be made publicly available upon acceptance of the paper, together with the code and the trained models.

According to the metrics defined in [20] [21], experimental results show that the proposed approach outperforms state-of-the-art light field disparity estimation methods for both densely and sparsely sampled LF. In addition, it does not require any prior information on disparity range as in [2], [4], [5] for example.

II. Related work

A. Stereo depth estimation

Depth estimation from stereo image pairs is a highly-studied vision problem. Scene depth is indeed needed for a variety of processing problems such as 3D reconstruction and view synthesis. The scene depth can be derived by computing the disparity between a stereo pair of views. As categorized in [22], most stereo algorithms consist of the following operations: matching cost computation, cost aggregation, disparity optimization and refinement. The matching cost measuring pixel dissimilarity can be based on the l_1 or l_2 norms computed within a fixed or adaptive window. The authors in [23] [24] use a *winner-takes-all* strategy to optimize the final disparity by choosing at each pixel the disparity associated with the minimal cost value. Other methods like graph cut [25] or coarse-to-fine refinement [26] are instead used for optimizing the final disparity. Besides the above methods computing cost volumes, methods based on statistical models, i.e. on Markov Random Field (MRF) [27] and Conditional Random Fields (CRFs) [28], have also been proposed.

While classical algorithms for extracting depth information from a rectified image pair compute pixel dissimilarity within a finite window as a matching cost, the authors in [29] train a CNN (convolution neural network) to predict similarity scores between two image patches, and compute the stereo matching cost. The authors in [30] propose a deep embedding model to map intensity values of image patches into an embedding feature space. Pixel dissimilarities are then measured by computing Euclidean distances between feature vectors. While the above methods compute matching costs on feature representations of rectified image pairs, the estimation problem still requires regularization or left-right consistency checks to have reliable estimates. The authors in [31] propose instead an end-to-end CNN framework with 3D convolutions to learn to regularize the cost volume as well as a soft argmin function to regress sub-pixel disparity values from the disparity cost volume.

In parallel of the above methods for estimating disparity between rectified image pairs, deep learning techniques have given momentum to a significant progress in optical flow estimation. The authors in [8] developed an end-to-end trainable encoder-decoder architecture with a correlation layer that explicitly provides matching capabilities between image pairs. Being a variant of *FlowNet*, instead of considering 2D correlation, *DispNet* [32] considers 1D correlation to better adapt to the disparity estimation task. The FlowNet network has been improved in *FlowNet 2.0* [9] by stacking several elementary networks similar to *FlowNet*. The structure of two parallel branches of sub-networks, one for large displacements prediction and another for small displacements, makes *FlowNet 2.0* applicable for variable flow ranges. In the continuity of their work, the authors in [10] perform a joint estimation of occlusions and optical flow in order to improve accuracy in occluded regions and along the contours. The authors in [11] construct a spatial pyramid using deep neural networks to learn the optical flow in a coarse-to-fine manner. A pyramid structure is also used in [12] to avoid computing a full cost volume that is computationally prohibitive for real-time optical flow estimation. Partial cost volumes are constructed by computing the distance between warped features of the second image and the features of the first image, within a limited search range, at each pyramid level. In [33], the authors propose a cascade network to refine the initial disparity estimation by learning, in a supervised fashion, residual signals across multiple scales.

B. Light field depth estimation

Different types of approaches have been proposed for scene depth estimation from light fields. A first category of methods derives the disparity by analyzing the epipolar plane images (EPI). Pixels in the different views corresponding to the same 3D point form a line in the EPI, whose slope is proportional to the disparity between the views [3]. The authors in [1] use structure tensors to locally estimate these slopes, this local estimation being then placed in a global optimization framework using a variational approach. The authors in [2] propose a spinning parallelogram operator for disparity estimation in the EPI, accompanied with a confidence measure to handle ambiguities and occlusions.

In contrast to EPI-based methods, the authors in [4] [5] [6] estimate disparity by searching for pixel matches between sub-aperture images (SAI). The authors in [4] estimate disparity by computing a matching cost volume between the central sub-aperture image and sub-aperture images warped using the phase shift theorem. The approach in [6] consists in estimating disparities between the four corner views, then propagating them to the target viewpoint. Correlation between viewpoints is exploited by a low rank approximation model to cope with occlusions. The authors in [5] employ an empirical Bayesian framework to estimate scene-dependent parameters for inferring scene disparity. This algorithm is free of additional cues exploiting dense view sampling, hence it is relevant for both dense and sparse light fields.

A deep learning architecture is proposed in [17] by introducing 3D convolutions on EPI volumes. The recently proposed EPINet approach [18] using a multistream approach achieves state-of-the-art performance. Each stream exploits one angular direction of light field views, horizontal, vertical, left or right diagonal directions. But these approaches are well suited for dense light fields only. The goal here is to design a neural architecture that would work well for both dense and sparse light fields.

III. Architecture overview

Let $L(x, y, u, v)$ denote a 4D representation of a light field, where (x, y) denote the spatial coordinates and (u, v) denote the angular coordinates. To simplify the notations, we will refer to a light field view by the index of its angular position, e.g., $L_{\mathbf{i}}$ where $\mathbf{i} = (u_i, v_i)$, and denote $d_{\mathbf{i}}^{\mathbf{j}}$ the disparity between two views $L_{\mathbf{i}}$ and $L_{\mathbf{j}}$ normalized by the distance between the two views.

The proposed learning framework to estimate depth (or disparity) for any light field viewpoint, from a subset of input views is depicted in Fig. 1. The approach is composed of three main steps, i.e. *stereo estimation*, *fusion* and *refinement*. We denote $L_{\mathbf{t}}$ the ***target view*** for which the disparity map is to be estimated. Multiple coarse disparity maps on this target position are first estimated by a convolutional network trained for stereo estimation. The model, that we call *FN2-ft-stereo*, is obtained by fine-tuning a pre-trained *FlowNet 2.0* network with light field stereo pairs (the details of this fine-tuning process will be explained in Section IV-A). Each of these disparity maps is computed between $L_{\mathbf{t}}$ (shadowed in blue) and a ***stereo view*** $L_{\mathbf{s}}$ located on the same row (framed in yellow) or on the same column (framed in red). For vertical image pairs, a rotation of 90 degrees is applied such that the displacement flow between these two images only contains the horizontal component. Accordingly, the obtained disparity map should be also rotated by 90 degrees in the reversed direction. The disparity maps between the target and the stereo views are denoted $d_{\mathbf{t}}^{\mathbf{s}}$, $\mathbf{s} \in S$, with S being the set of used stereo view positions.

These multiple estimates of the disparity information on the target view $L_{\mathbf{t}}$ should be fused to a single disparity map. To achieve this, we leverage the warping error from a set of ***anchor views*** (framed in blue) $L_{\mathbf{a}}$, $\mathbf{a} \in A$, with A denoting the set of anchor view positions. The disparity value corresponding to the smallest error per pixel is selected for the fused disparity map. In order to better handle the object boundary, the warping error is computed differently for pixels within occlusion areas or those within occlusion-free areas.

This fusion is simple and efficient, but is prone to noise and discontinuity because the decision is made pixel by pixel. Further refinement is realized by a second CNN which learns in a supervised fashion the residual signals of the disparity at multiple scales by an encoder-decoder architecture. Our network structure is flexible with respect to the anchor views, i.e. anchor views can be located at any viewpoint of the light field, and no additional training is required if the anchor view positions are changed.

IV. Processing workflow

A. Fine-tuned FlowNet 2.0 for disparity estimation

FlowNet 2.0 (*FN2*) [9] is an efficient CNN-based optical flow estimator. Two parallel branches of sub-networks are combined, the first specialized in large displacements estimation and the second in small displacements. The final stage of the network merges the two previously estimated flows taking into account the flow magnitude. Thanks to this structure, it is relevant to apply *FN2* to estimate disparity for light field views with variable disparity ranges.

Let us denote $f(L_{\mathbf{i}}, L_{\mathbf{j}}) = (f^x_{\mathbf{i}\to\mathbf{j}}, f^y_{\mathbf{i}\to\mathbf{j}})^\top$ the flow estimation operator between the views $L_{\mathbf{i}}$ and $L_{\mathbf{j}}$. Assuming that the light field is well rectified and regularly spaced in both angular directions, the disparity map between the view $L_{\mathbf{i}}$ and $L_{\mathbf{j}}$, normalized by the distance between the views, can be computed as

$$d_{\mathbf{i}}^{\mathbf{j}} = \frac{f^x_{\mathbf{i}\to\mathbf{j}}}{v_i - v_j} = \frac{f^y_{\mathbf{i}\to\mathbf{j}}}{u_i - u_j}. \tag{1}$$

In order to well adapt the operator $f(\cdot,\cdot)$ to disparity estimation between two light field views, we fine-tune the pre-trained *FN2* model. Two strategies have been considered. The first one feeds the model with pairs of light field views $L_{\mathbf{i}}$ and $L_{\mathbf{j}}$ with no constraint on view positions, and the model learns dense optical flows both on the horizontal and vertical directions. The obtained model is denoted *FN2-ft*. Another approach is to learn the model using image pairs $L_{\mathbf{i}}$ and $L_{\mathbf{j}}$ on the same row ($u_i = u_j$) or on the same column ($v_i = v_j$). Note that images on the same column are rotated 90 degrees to become a horizontal stereo pair $(rot(L_{\mathbf{i}}), rot(L_{\mathbf{j}}))$. The obtained model is thus named *FN2-ft-stereo*. Formally, with *FN2-ft-stereo*, the disparity map for the view $L_{\mathbf{i}}$ is computed as

$$d_{\mathbf{i}}^{\mathbf{j}} = \begin{cases} \frac{f^x_{\mathbf{i}\to\mathbf{j}}}{v_i - v_j}, & \text{if } u_i = u_j \\ \frac{rot^{-1}(f^x_{\mathbf{i}^*\to\mathbf{j}^*})}{u_i - u_j}, & \text{if } v_i = v_j \end{cases} \tag{2}$$

where $f^x_{\mathbf{i}^*\to\mathbf{j}^*}$ denotes the horizontal flow component between $L_{\mathbf{i}^*}$ and $L_{\mathbf{j}^*}$ with $L_{\mathbf{i}^*} = rot(L_{\mathbf{i}})$. The symbols $rot(\cdot)$ and $rot^{-1}(\cdot)$ are counterclockwise and clockwise rotation of $90°$.

In Fig. 2, we compare the estimation accuracy of the *FN2-ft-stereo* model against three other models: *FN2* (the pre-trained *FlowNet 2.0* model), *FN2-ft* and *DispNet-ft*, which is

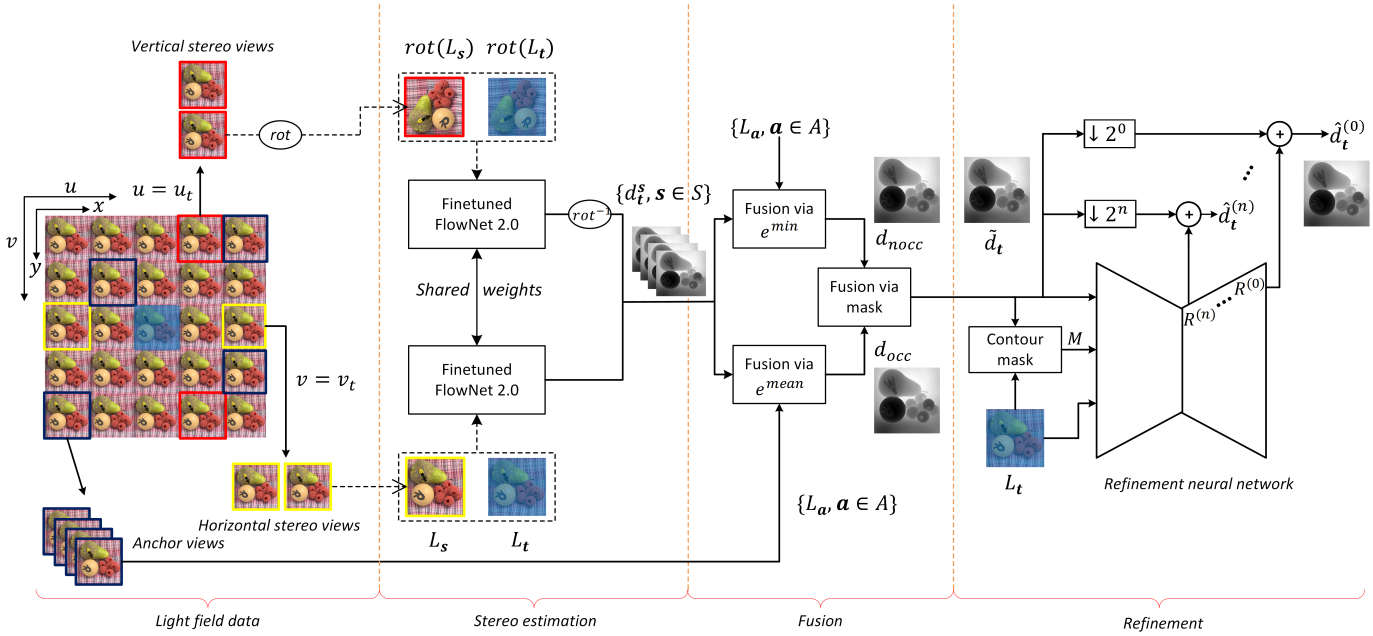


Fig. 1. Overview of proposed framework. We take a 5×5 LF as example. The blue masked view, called ***target view*** $L_{\mathbf{t}}$, is the view for which we search to estimate the disparity. Views in the yellow and red rectangles are respectively horizontal and vertical ***stereo views*** denoted $L_{\mathbf{s}}$. Target and stereo views are used to compute the initial disparity maps d_i using a fine-tuned *FlowNet 2.0* model. ***Anchor views*** $L_{\mathbf{a}}$ (in dark blue rectangles) can be any subset of views, except the target view, that are used to compute the warping error for the fusion of initial estimates. A multi-scale residual learning network corrects fusion artifacts and smoothes the final disparity map in a last refinement step.

obtained by finetuning a pre-trained *DispNet* model [32] with our stereo light field views. The models *FN2* and *FN2-ft* can estimate displacement both in x and y dimensions, whereas *DispNet-ft* and *FN2-ft-stereo* focus on 1D (horizontal or vertical displacements) estimation. On one hand, *FN2-ft-stereo* performs better than *FN2* and *FN2-ft*, which shows the necessity of concentrating on 1D estimation. In addition, *FN2-ft-stereo* is significantly better than *DispNet-ft*, both being finetuned using the same training set of stereo light field views.

Therefore, we choose to use *FN2-ft-stereo* for computing a set $D_{\mathbf{t}}$

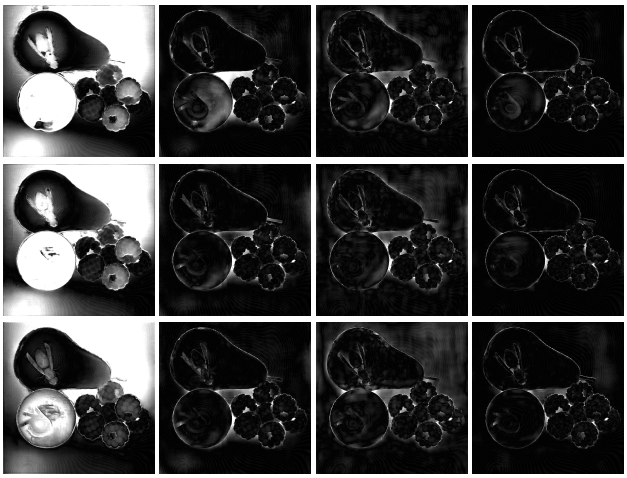
$$D_{\mathbf{t}} = \{d_{\mathbf{t}}^{\mathbf{s}}, \mathbf{s} \in S\} \tag{3}$$

of multiple estimates of disparity $d_{\mathbf{t}}^{\mathbf{s}}$ between the target view $L_{\mathbf{t}}$ and one of the stereo views $L_{\mathbf{s}}$. As each of the candidates $d_{\mathbf{t}}^{\mathbf{s}}$ is normalized by the distance between the views in the considered pair, it represents the amount of disparity between the view and its immediate neighboring views. In the sequel, we will denote this set of normalized disparity maps as $D_{\mathbf{t}} = \{d_k, k = 1..K\}$, with K the number of candidate maps.

B. Fusion based on warping error maps

Although our *FN2-ft-stereo* model provides satisfying results for disparity estimation with stereo pairs, information in other available views of the light field is not exploited. In this subsection, we propose to fuse the candidate maps in $D_{\mathbf{t}}$ into a single disparity map based on the error of warping the anchor views $L_{\mathbf{a}}$, $\mathbf{a} \in A$ onto the target view.

Based on one of the disparity maps $d_k \in D_{\mathbf{t}}$, backward warping is applied to project the anchor view $L_{\mathbf{a}}$ to the target position $\mathbf{t}$. The warped view is denoted $\widetilde{L}^k_{\mathbf{a}\to\mathbf{t}}$. The



(a) *FN2* (b) *FN2-ft* (c) *DispNet-ft* (d) *FN2-ft-stereo*

Fig. 2. Disparity estimation errors (display range between 0 and 1) using different models: (a) *FN2*, (b) *FN2-ft*, (c) *DispNet-ft* and (d) *FN2-ft-stereo*. The first row corresponds to the estimation errors using a stereo pair $L_{2,2}$ and $L_{2,8}$. On the second and third row, for *FN2* and *FN2-ft*, the estimation has been done between the views $L_{5,5}$ and $L_{8,8}$, and the horizontal (second row) and vertical (third row) flow components are shown. Since *DispNet-ft* and *FN2-ft-stereo* only take stereo pairs, the horizontal flows are estimated between $L_{5,5}$ and $L_{5,8}$ (the second row), and the vertical flows are estimated between $L_{5,5}$ and $L_{8,5}$ (the third row).

corresponding warping error $e_k^{\mathbf{a}}$ is computed by summing on the three R, G, B color channels:

$$\forall \mathbf{a} \in A, e_k^{\mathbf{a}} = \sum_{R,G,B} (L_{\mathbf{t}} - \widetilde{L}^k_{\mathbf{a}\to\mathbf{t}})^2 \tag{4}$$

Warping errors are then fused by taking into account all the warped views $\widetilde{L}^k_{\mathbf{a}\to\mathbf{t}}$ with $\mathbf{a} \in A$. The fusion is performed

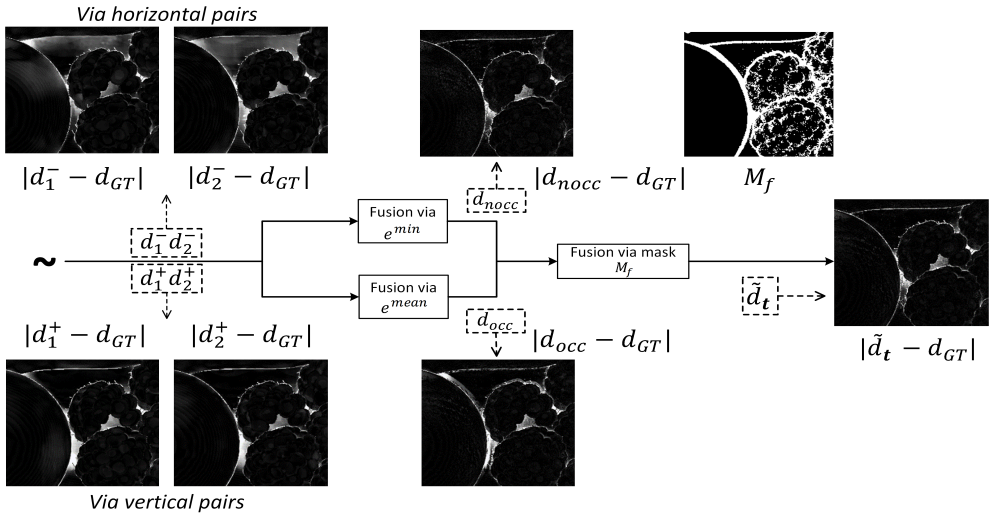


Fig. 3. Disparity error (compared to the ground truth) at each step of the fusion process. d_1^-, d_2^- and d_1^+, d_2^+ are coarse disparity estimates using *FN2-ft-stereo*, computed with horizontal or vertical image pairs ('-' indicates horizontal image pairs and '+' indicates vertical ones). d_{nocc} and d_{occ} are respectively fused disparity maps via e^{mean} and e^{min}, and $\tilde{d}$ is the final resulting map using the binary mask M_f.

either by the "average" or by "min" operations as

$$e_k^{\text{mean}}(\mathbf{p}) = \text{mean}_{\mathbf{a}} e_k^{\mathbf{a}}(\mathbf{p}), \mathbf{a} \in A \tag{5}$$

$$e_k^{\text{min}}(\mathbf{p}) = \text{min}_{\mathbf{a}} e_k^{\mathbf{a}}(\mathbf{p}), \mathbf{a} \in A. \tag{6}$$

Both the error maps e_k^{mean} and e_k^{min} suggest the reliability on values in the disparity map d_k, but possess complementary properties. The error map e_k^{mean} reflects well the disparity inaccuracy in the occlusion-free zones, since it averages the contribution from all the warped views. Nevertheless, in occluded areas, interpolation in large holes becomes the main cause of warping errors instead of disparity inaccuracy. In this case, e_k^{min} turns out to be a more relevant measure. Indeed, at a pixel $\mathbf{p}$ that can be seen in the warped view $\widetilde{L}_{\mathbf{a}'}^k$, but not in another warped view $\widetilde{L}_{\mathbf{a}}^k, \mathbf{a} \in A, \mathbf{a} \neq \mathbf{a}'$, the value $e_k^{\text{mean}}(\mathbf{p})$ is misleading because of the high contribution of the error $e_k^{\mathbf{a}}(\mathbf{p})$. On the contrary, the "min" operation gets rid of the perturbation from the occluded views. However, if a pixel $\mathbf{p}$ is occluded in all the warped views, neither $e_k^{\text{mean}}(\mathbf{p})$ nor $e_k^{\text{min}}(\mathbf{p})$ gives a reliable measure of disparity inaccuracy. It is preferable that the anchor view positions $\mathbf{a}$ are dispersed in the light field such that a pixel occluded in one view may be seen in another view. The impact of anchor view positions on the quality of the final disparity map will be discussed in Section VII-B.

To fuse at each pixel $\mathbf{p}$ the candidate disparity values $d_k(\mathbf{p}), k = 1..K$, a *winner-takes-all* strategy is employed according to error values $e_k^{\text{mean}}(\mathbf{p})$ and $e_k^{\text{min}}(\mathbf{p})$:

$$k' = \arg\min_k e_k^{\text{min}}(\mathbf{p}) \tag{7}$$

$$d_{\text{occ}}(\mathbf{p}) = d_{k'}(\mathbf{p}) \tag{8}$$

and

$$k'' = \arg\min_k e_k^{\text{mean}}(\mathbf{p}) \tag{9}$$

$$d_{\text{nocc}}(\mathbf{p}) = d_{k''}(\mathbf{p}) \tag{10}$$

Two fused disparity maps are obtained, d_{nocc} for occlusion-free zones and d_{occ} for occluded areas. To reduce local inconsistency, a 3×3 mean filter is applied on the error maps e_k^{mean} and e_k^{min}. A binary mask M_f defined as

$$M_f(\mathbf{p}) = \begin{cases} 1 & \text{min}_k(e_k^{\text{mean}}(\mathbf{p})) > \theta \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

is used to merge these two resulting disparity maps. The value $M_f(\mathbf{p})$ equals to 1 if $\text{min}_k(e_k^{\text{mean}}(\mathbf{p}))$ exceeds a certain scene-dependent threshold θ, which is fixed at the value at the 90 percentile (errors at the occluded pixels are generally higher than those at the non-occluded ones). Note that M_f is not the real occlusion mask contrary to that used in [6]. However, it can be computed much more efficiently, and it approximates well the real mask.

Finally, one unique disparity map at the target position $\mathbf{t}$ is obtained as

$$\widetilde{d}_{\mathbf{t}} = d_{\text{nocc}} \odot M_f + d_{\text{occ}} \odot (1 - M_f) \tag{12}$$

where $\odot$ denotes pixel-wise Hadamard product. Fig. 3 demonstrates the gain in the fusion process. The errors on the estimated disparity compared to the ground truth are illustrated for each step of the process.

C. Multi-scale disparity refinement

The fusion step enables us to take advantage of multiple estimates and significantly improves the estimation accuracy. Nevertheless, as the fusion is implemented by a per pixel *winner-takes-all* (WTA) selection, discontinuity may exist in the resulting disparity map. Further refinement operation is useful for quality enhancement.

The work in [33] proposed a two-stage disparity learning framework. The learned disparity between a stereo pair is refined with a multi-scale residual learning network. As input, their network takes two stereo color images, the previously

estimated disparity map, the warped image, as well as the corresponding warping error image. In a multi-view stereo scenario where we exploit the color information from multiple stereo views $L_{\mathbf{s}}, \mathbf{s} \in S$ and anchor views $L_{\mathbf{a}}, \mathbf{a} \in A$, it is obvious that this structure is no longer applicable. Indeed, the large number of input views, as well as the number of initialized disparity maps, will rapidly enlarge the size of the network and increase the computational cost during training. Moreover, the scheme is not flexible with respect to the varying number of input views.

The fusion of disparity estimates (Section IV-B) partially resolves this problem. Regardless of the variable number of stereo views, as well as that of the initialized disparity maps, only one single disparity map $\widetilde{d}_{\mathbf{t}}$ has been obtained for the target position $\mathbf{t}$. Besides $\widetilde{d}_{\mathbf{t}}$, two other images are fed to our refinement network: the target view $L_{\mathbf{t}}$ serving as color guidance and a binary mask M indicating contour misalignment. The mask M is computed as

$$M = |\Gamma(\Psi(\widetilde{d}_{\mathbf{t}})) - \Gamma(\Psi(\widetilde{d}_{\mathbf{t}})) \odot \Gamma(\Psi(L_{\mathbf{t}}))| \tag{13}$$

with $\Gamma(\cdot)$ being the dilation operator and $\Psi(\cdot)$ being the canny contour detector.

Compared to the network used in [33], we change the inputs and the first layer of the refinement network, and we construct a 9-layer convolutional encoder to extract features and a 16-layer decoder to retrieve the estimated disparity map. The proposed network contains about 3.6×10^7 trainable parameters. During training, the residual signals of the disparity are learned at different resolution scales $n \in [0, N]$, supervised by the ground truth disparity. At each scale n, the network generates the residual signal $R^{(n)}$, which is added to the downsampled disparity map $\widetilde{d}_{\mathbf{t}}^{(n)}$

$$\widehat{d}_{\mathbf{t}}^{(n)} = \widetilde{d}_{\mathbf{t}}^{(n)} + R^{(n)}, \tag{14}$$

where $\widehat{d}_{\mathbf{t}}^{(0)}$ denotes the final obtained full resolution disparity map.

The loss is summed over all the resolution scales $n \in [0, N]$ as

$$\mathcal{L} = \sum_{n=0}^{N} \mu_n \mathcal{L}^{(n)}, \tag{15}$$

where μ_n is the contribution weight for the loss at the scale n and $\mathcal{L}^{(n)}$ is the sum of two losses

$$\mathcal{L}^{(n)} = \lambda_1 \mathcal{N}(\widehat{d}_{\mathbf{t}}^{(n)}, d_{\mathrm{GT}}^{(n)}) + \lambda_2 \mathcal{G}(\widehat{d}_{\mathbf{t}}^{(n)}, d_{\mathrm{GT}}^{(n)}), \tag{16}$$

where $\mathcal{N}$ denotes the sum of absolute differences (SAD)

$$\mathcal{N}(d_1, d_2) = \sum_{\mathbf{p}} |d_1(\mathbf{p}) - d_2(\mathbf{p})| \tag{17}$$

and where $\mathcal{G}$ is a gradient term defined as

$$\mathcal{G}(d_1, d_2) = \sum_{\mathbf{p}} \|G(d_1, d_2, \mathbf{p})\|_2 \tag{18}$$

with

$$G(d1, d2, \mathbf{p}) = \Big(\nabla_x d_1(\mathbf{p}) - \nabla_x d_2(\mathbf{p}), \nabla_y d_1(\mathbf{p}) - \nabla_y d_2(\mathbf{p})\Big)^{\top}. \tag{19}$$

V. Datasets

The effectiveness of data-driven algorithms significantly depends on the quality and the quantity of training data. Supervised learning of neural models for depth or disparity estimation [32] [17] requires large datasets with ground truth disparity information. A few datasets of synthetic light fields are publicly available. The MIT Light Field Archive [34] includes 17 light fields with angular resolution of 5×5 or 7×7 views, but the ground truth disparity maps are not provided. Two HCI synthetic light field datasets exist. The dataset [35] contains 8 light fields with disparity information for all the views, each light field containing 9×9 views of 768×768 pixels. Recently, a second dataset [20] is released containing 24 light field scenes with a spatial resolution of 512×512 and an angular resolution of 9×9. Among them, 16 scenes are provided with disparity maps for all the views, whereas for the 8 others the disparity information is available only for the central view. In addition, these datasets are limited to densely sampled light fields with narrow baselines.

Since our goal is to propose a framework applicable to both densely and sparsely sampled light fields, we have created two synthetic datasets: a Sparse Light Field Dataset (SLFD) including 53 scenes with disparity range $[-20, 20]$ pixels between adjacent views, and a Dense Light Field Dataset (DLFD) containing 43 scenes with disparity range $[-4, 4]$ pixels. Each light field has the same resolution $512 \times 512 \times 9 \times 9$ as those in the HCI dataset [20]. Both SLFD and DLFD are provided with the disparity and the depth maps for every viewpoint in the light fields. To the best of our knowledge, SLFD is the first sparse synthetic light field dataset which provides ground truth depth and disparity information for every light field view.

The rendering of the light field scenes is performed with the open source software Blender [36]. The elementary models are downloaded from the websites Chocofur [37] and Sketchfab [38] with a non-commercial CC license, and are assembled to create various 3D, mostly indoor scenes. The scenes contain textureless background, specular reflection, diffusion and object occlusion, which makes our dataset useful to measure the effectiveness of depth estimation algorithms. The 3D scene models in SLFD and DLFD are partly shared, but they are rendered with different camera baselines.

The dataset SLFD is split into a training set of 44 scenes and a valid set of 9 scenes, whereas DLFD is split into a training set of 38 scenes and a test set of 5 scenes. Fig. 4 shows some examples of light field scenes and their corresponding disparity maps. For training the network, we have also used 16 scenes of the HCI 4D light field benchmark dataset [20] together with our DLFD.

VI. Implementation details

A. Training data preparation

For fine-tuning *FlowNet 2.0*, stereo views are extracted on the same row or the same column of a light field. Image pairs located on the same column are rotated with a counterclockwise 90° to convert vertical pixel displacement to horizontal displacement. The two images in an extracted

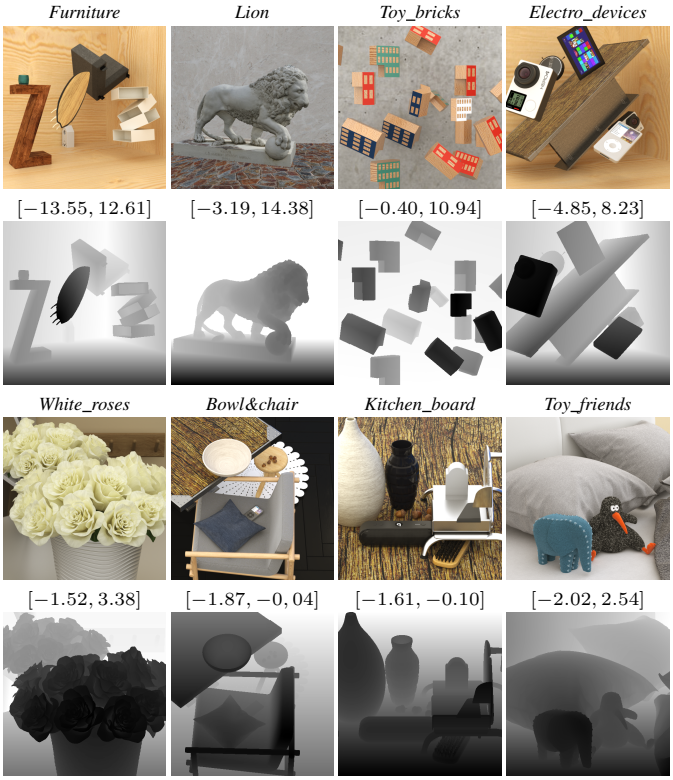


Fig. 4. Examples of scenes from our two datasets, the 1st and 2nd rows show three scenes and the corresponding disparity maps from SLFD, while the 3rd and the 4th rows show three examples of scenes and the corresponding disparity maps from DLFD.

pair are separated by an angular *distance* corresponding to a view index difference $l \in [2, 3, ...8]$ for dense light fields, which corresponds to a disparity range within [-32,32] pixels, whereas for sparse light fields, this *distance* is set to be $l \in [1, 2, 3]$, corresponding to a disparity range within [-60,60] pixels. In both cases, the extraction of views is done in such a way that the different distances (or disparities) are well represented (with same probability) in the training data.

B. Data augmentation

The authors in [8] [17] performed geometrical and chromatic transformations to increase diversity in the training data. In our experiments, however, we have found that geometrical transformations such as rotation, translation or scaling that involve data interpolation bring extra errors in the ground truth disparity values, and thus harm the learning convergence. As a consequence, only chromatic transformation has been applied by changing the hue, saturation, contrast and brightness of training images. Concretely, we convert the images from the RGB space to the HSV space, add an offset to the hue and saturation channels, and then convert the images back to RGB color space. The hue and saturation offsets are uniformly picked from $[-0.3, 0.3]$ and $[0.7, 1.3]$. To perform contrast augmentation, we compute the mean pixel values $\bar{c}$ of each image channel c, then adjust c to $(c - \bar{c}) \times \zeta + \bar{c}$, where ζ is a contrast factor uniformly picked from $[0.7, 1.3]$. The brightness augmentation is implemented by adding a brightness offset to each of the RGB channels of an image, which is randomly picked from $[-0.1, 0.1]$.

C. Learning details

Different learning schedules are employed for fine-tuning the *FN2-ft-stereo* model and for training the refinement network. In the finetuning step, thanks to the pre-trained model, a shorter learning schedule can be adopted with an initial learning rate set to 0.0001 for the first 500 epochs. The learning rate is then decreased by half every 200 epochs. For the training of the refinement network which is randomly initialized, the schedule is longer with an initial learning rate of 0.0001 for the first 1200 epochs. The learning rate is then divided by 2 every 200 epochs. We use the Adam optimizer [39], and becaused of the limited GPU memory, a batch size of 4 is used. Tensorflow [40] is used to implement our network. It takes about 2 days to train our network with a NVIDIA Tesla P100 GPU with 16G memory.

VII. Experimental results

A. Setup

To validate the effectiveness of our proposed framework, we conduct experiments on both public and self-rendered synthetic datasets and with real light fields.

1) Synthetic Dataset: For sake of comparison, we use the synthetic light fields of the HCI datasets [20] [35] and keep the same test light fields as in [6]: *Stilllife, Buddha, Butterfly, MonasRoom* from [35] and *Boxes, Cotton, Dino, Sideboard* from [20]. The 12 additional scenes of [20] are added in the training set as detailed in Section V.

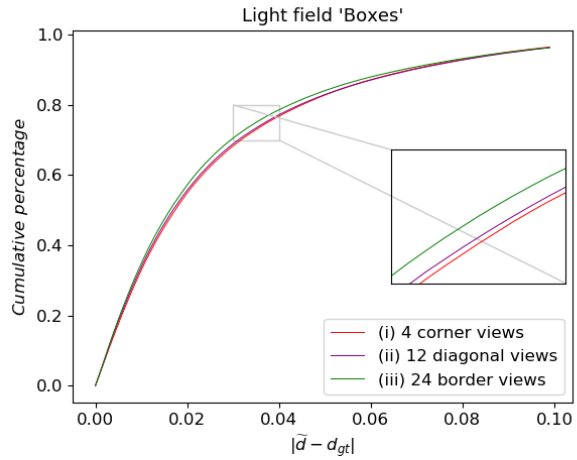
We also evaluated the proposed framework using our own sparse light fields datasets (that will be made publicly available at the time of the paper publication). Four test light fields *Furniture, Lion, Toy_bricks, Electro_devices* are used for evaluation. The scene *Lion* contains a single object and the other three scenes contain multiple objects.

2) Real Light Fields: We have also tested our framework with dense real light fields, using datasets captured by plenoptic Lytro Illum cameras (we used light fields in the INRIA [41] and EPFL [42] datasets). Compared with synthetic datasets, light fields captured by plenoptic cameras are more challenging due to the fact that the extracted views contain noise and geometrical distortions. These real light fields have a spatial resolution of 434×625 pixels and an angular resolution of 15×15 views. Finally, experiments have been also carried out for sparse real world light fields captured with wide baseline camera arrays [43].

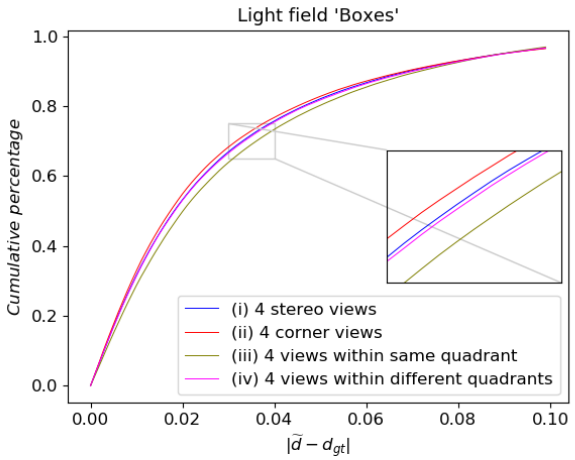
B. Impact of the anchors views

In contrary to other deep learning frameworks [17], [18], our network is flexible with respect to the number and the positions of the input views. Indeed, it is possible to arbitrarily select a subset of light field views as anchor views.

Fig. 5 evaluates the percentage of pixels below a certain error threshold for different strategies to select anchor views. The higher is this percentage, more accurate is the estimation.



(a) Varying the number of anchor views:(i) 4 corner views; (ii) 12 views on the diagonals; (iii) 24 border views.



(b) Anchor views at different positions: (i) 4 anchor views = 4 stereo views; (ii) 4 anchor views = 4 corner views; (iii) 4 views located in the same quadrant; (iv) 4 views randomly selected such that each quadrant contains one anchor view.

Fig. 5. Percentage of pixels below a certain error threshold for different strategies to select anchor views: (a) varying the number of anchor views; (b) with a fixed number (4) of anchor views at different positions.

TABLE I
QUANTITATIVE COMPARISON WITH NON-LEARNING-BASED METHODS ON SYNTHETIC LIGHT FIELDS DATASETS

Light fields	MSE					BP1					BP2					BP3				
	[4]	[2]	[5]	[6]	Ours	[4]	[2]	[5]	[6]	Ours	[4]	[2]	[5]	[6]	Ours	[4]	[2]	[5]	[6]	Ours
Stilllife	2.02	1.72	1.53	2.56	**1.07**	81.2	76.2	76.0	71.3	**70.5**	51.0	32.1	41.0	25.0	**24.8**	20.9	6.8	16.2	9.2	**5.8**
Buddha	1.13	0.97	0.49	0.82	**0.41**	57.7	41.2	52.6	**34.9**	35.3	24.4	14.8	15.0	12.3	**7.7**	10.1	6.7	2.9	5.4	**2.2**
MonasRoom	0.76	0.58	0.66	0.53	**0.39**	46.0	42.5	48.2	**38.6**	39.0	22.1	17.8	20.2	18.6	**13.7**	11.7	7.8	10.4	8.2	**6.1**
Butterfly	4.79	0.74	0.80	1.84	**0.58**	82.5	78.9	83.4	**70.8**	73.4	49.1	48.5	50.9	**36.0**	42.0	15.4	14.1	17.6	6.7	**6.3**
Boxes	14.15	**8.23**	11.30	12.71	9.16	72.7	**62.3**	87.2	65.8	68.4	45.5	**28.1**	65.0	37.7	38.5	26.4	**15.8**	42.0	23.9	22.1
Cotton	9.98	1.44	2.04	1.18	**0.94**	60.5	41.7	75.8	42.6	**38.2**	23.3	11.1	37.5	10.7	**10.6**	8.9	**2.7**	10.4	4.1	3.3
Dino	1.23	**0.29**	0.67	0.88	0.50	76.6	57.5	84.8	49.1	**45.6**	48.4	17.9	57.2	20.0	**14.5**	20.9	**3.4**	24.1	9.5	4.7
Sideboard	4.16	**0.92**	1.34	10.31	1.37	67.8	64.3	78.6	**61.7**	63.6	39.3	31.0	44.1	37.5	**26.3**	23.0	10.4	15.0	19.6	**10.1**
Average	4.78	1.86	2.35	3.85	**1.80**	68.1	58.1	73.3	54.4	**54.3**	37.9	25.2	41.4	24.7	**22.3**	17.2	8.5	17.3	12.1	**7.6**
Furniture	-	-	**0.37**	1.94	0.39	-	-	86.3	41.3	**40.7**	-	-	73.1	41.3	**23.0**	-	-	36.0	20.2	**8.6**
Lion	-	-	0.10	0.87	**0.09**	-	-	**35.9**	73.0	47.6	-	-	23.9	59.5	**9.0**	-	-	5.5	9.5	**2.6**
Toy_bricks	-	-	**0.22**	1.10	0.56	-	-	59.5	66.4	**50.5**	-	-	33.2	44.6	**23.7**	-	-	**4.7**	11.2	12.4
Electro_devices	-	-	0.20	0.63	**0.19**	-	-	76.9	60.7	**52.8**	-	-	57.4	43.4	**30.5**	-	-	22.5	18.6	**8.7**
Average	-	-	**0.22**	1.14	0.31	-	-	64.7	64.7	**47.9**	-	-	46.9	47.2	**21.6**	-	-	17.2	14.9	**8.1**

*Number of input views: [4]-**49 views**, [2]-**49 views**, [5]-**5 views**, [6]-**4 views**, Ours-**5 views**
*For the first 8 scenes (dense LFs), MSE denotes 100*Mean Square Error, BP1, BP2, BP3 denote Bad Pixel Ratios with thresholds 0.01, 0.03, 0.07.
*For the last 4 scenes (sparse LFs), MSE denotes Mean Square Error, BP1, BP2, BP3 denote Bad Pixel Ratios with thresholds 0.05, 0.1, 0.3.

We consider the 7×7 central views of the light field "Boxes" is studied. Fig. 5(a) assesses the impact of the number of anchor views on the final estimation accuracy. The 4 corner views, the 12 views on the diagonals and the 24 views on the border of the light field are respectively used as the subset A of anchor views. We observe that using more anchor views is useful for improving estimation accuracy, though the improvement may be limited. This suggests that when the time consumption or GPU memory is the bottleneck, less anchor views can be exploited without too much degrading the estimation accuracy.

In Fig. 5(b), the number of anchor views is fixed to 4, and we evaluate the impact of the anchor view positions on the performance. The assessed sets A of the anchor views can be: (i) 4 stereo views ($A = S$), (ii) 4 corner views, (iii) 4 views located in the same quadrant (the light field can be divided into four quadrants, "northeast", "southeast", "southwest" and "northwest", according to the location with respect to the target view), and (iv) 4 views randomly selected such that each quadrant contains one anchor view. The strategy (iii) achieves the worst performance, since the 4 views located in the same quadrant do not contain occlusion information of the other quadrants. The use of the stereo views as anchor views (i) obtains worse performance than (ii), since the geometry information of the stereo views is already exploited in the coarse estimation step. And indeed, for a dense light field such as "Boxes", geometry information of the 3D scene can be mostly recovered from the 4 corner viewpoints.

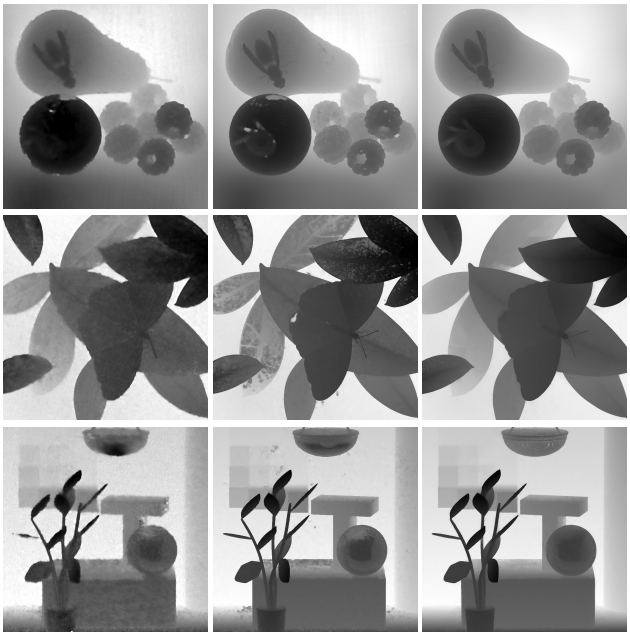
(a) *Jeon et al.* [4] (b) *Zhang et al.* [2] (c) *Huang* [5] (d) *Jiang et al.* [6] (e) *Ours* (f) *GT* (g) *Image*

Fig. 6. Qualitative comparison with non deep learning-based methods. Each row shows the estimated disparity maps with two zoomed areas (homogeneous area framed in red and contour area framed in blue) for different methods: (a) *Jeon et al.* [4], (b) *Zhang et al.* [2], (c) *Huang* [5], (d) *Jiang et al.* [6], (e) *Our framework*. The (f) *Ground truth* and (g) *Color image* are also shown.

TABLE II
QUANTITATIVE COMPARISON WITH LEARNING-BASED METHODS ON SYNTHETIC LIGHT FIELDS DATASETS

Light fields	MSE			BP1			BP2			BP3		
	[17]	[18]	Ours	[17]	[18]	Ours	[17]	[18]	Ours	[17]	[18]	Ours
Stilllife	3.02	1.96	**1.14**	84.2	77.3	**71.0**	56.2	39.4	**25.8**	22.9	11.5	**6.2**
Buddha	0.52	**0.26**	0.43	75.9	39.9	**35.7**	37.3	**5.2**	8.0	9.1	**1.4**	2.3
MonasRoom	1.06	0.60	**0.41**	76.8	42.5	**39.6**	41.9	14.5	**14.3**	16.2	7.8	**6.4**
Butterfly	1.13	1.41	**0.57**	85.5	84.3	**72.4**	57.2	59.7	**40.2**	22.4	24.1	**6.0**
Boxes	9.06	**5.20**	9.97	82.5	**62.4**	68.4	53.7	**27.4**	39.6	30.5	**15.1**	23.6
Cotton	0.97	**0.25**	0.76	77.8	51.7	**36.6**	42.0	**4.9**	10.0	10.9	**0.9**	2.9
Dino	1.25	**0.19**	0.53	83.6	**41.0**	45.1	54.4	**6.6**	14.6	23.7	**1.9**	5.0
Sideboard	2.33	**0.80**	1.45	82.9	**58.6**	66.0	54.1	**21.0**	27.9	24.6	**6.6**	10.9
Average	2.42	**1.33**	1.91	81.2	57.2	**54.4**	49.6	**22.3**	22.6	20.0	8.7	**7.9**
Furniture	9.18	1.73	**0.42**	96.4	85.1	**40.2**	92.8	71.3	**23.0**	78.8	38.4	**8.9**
Lion	1.59	3.41	**0.09**	95.3	87.4	**48.7**	90.6	76.4	**8.7**	72.9	56.3	**2.6**
Toy_bricks	3.70	**0.36**	0.57	96.0	85.1	**49.5**	92.0	70.6	**23.3**	76.5	29.6	**12.6**
Electro_dev	7.82	0.74	**0.20**	95.0	80.3	**51.5**	89.9	60.6	**29.2**	72.0	22.8	**8.9**
Average	5.57	1.56	**0.32**	95.7	84.5	**47.5**	91.3	69.7	**21.1**	75.0	36.8	**8.3**

*Number of input views: [17]-**7 views (dense), 3 views (sparse)**, [18]-**25 views (dense), 9 views (sparse)**, Ours-**5 views**, MSE and BP are kept as same as those in Table I.



(a) *Heber et al.* [17] (b) *Shin et al.* [18] (c) *Ours*

Fig. 7. Qualitative comparison to deep learning-based methods, with methods (a) *Heber et al.* [17], (b) *Shin et al.* [18], (c) *Ours*.

C. Results with Densely Sampled Synthetic Light Fields

We compare our approach with both traditional and learning based state-of-the-art methods for densely sampled light fields. First, 4 reference methods [2] [4] [5] [6] using traditional approaches are considered. The disparity range is discretized for the methods [2], [4], [5]. As suggested in the light field depth estimation challenge held in 2017 LF4CV workshop [21], the number of disparity levels is set to 100 for the method [4] and 256 for the method [2]. For the method [5], the disparity step is set to 0.01, which corresponds to the minimal threshold of bad pixel ratios that we use. Both explicit and implicit discretization operations in [4] [2] [5] need disparity ranges as priors. To estimate the disparity map for the central view, the methods [2] [4] exploit the whole light field containing 49 views. The method in [6] takes four corner views to infer the central view disparity while the method in [5] chooses 5 images in the crosshair with target view in the center. For our framework, we employ the same crosshair pattern as [5] with 4 images serving as stereo and anchor views at the same time.

The upper part of Table I compares the estimation accuracy obtained with 8 HCI test scenes using different metrics: Mean Square Error (MSE) and Bad Pixel Ratios (BP) with thresholds 0.01, 0.03 and 0.07 (BP represents the percentage of pixels having an error superior to a certain threshold). In the experiment, we consider the central 7×7 sub-aperture images of the light field and estimate the disparity of the central view. The experiment shows that our framework achieves superior performances compared with other methods for most of the scenes both in terms of MSE and BP. In some cases, our framework yields the second best results with a slight difference only with the method ranked first. Compared with the methods in [2] [4] which exploits all the light field views, our method gives better results in spite of the fact that we use only a subset of light field views. In comparison with the other two methods [5] [6] using a subset of light field views, our method is competitive and sometimes wins with a large margin.

Fig. 6 shows the estimated disparity maps for the central view. Readers are recommended to zoom and view these results on the screen for visual comparison. The methods in [4] and [6] obtain disparity maps with distorted boundaries, while the method in [5] loses precision on slanted surfaces where disparity values gradually vary. In contrast, the methods [2] [5] as well as our framework can estimate disparities with more precise boundaries. Our method, although it may suffer from a subtle smoothness along boundaries, it yields less artifacts within homogeneous and slanted areas, and gives more visually pleasing results.

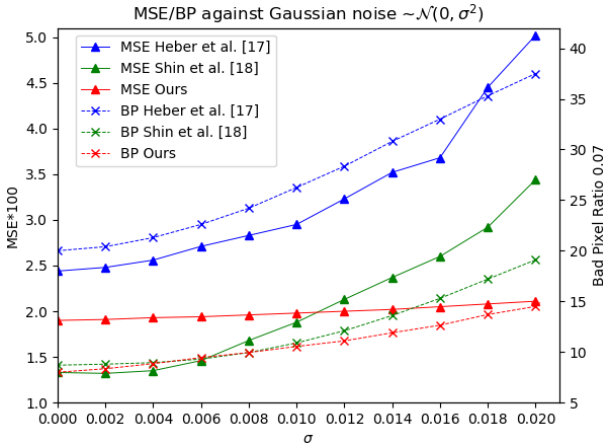
We also compare our method against two state-of-the-art methods based on deep learning [17], [18]. Both of them exploit epipolar geometry of the light field. The method in [17] retrieves disparity values by exploiting 3D EPI volumes containing texture information on two spatial dimensions and one angular dimension, whereas the method in [18] constructs the input volume by taking views on four different angular directions: horizontal, vertical, left and right diagonals. As these trained models are not publicly available for 7×7 light field views, we re-trained the models following the instructions in the corresponding papers. Since the network in [18] was implemented without zero-padding after each convolutional layer, the resulting disparity map loses 11 pixels at each border. For the sake of comparison, we cropped the same amount of bordering pixels for the method in [17] and our method as well (this explains the slight difference of measurement of our method in Table I and Table II). Quantitative results are shown in Table II. Our proposed method has several advantages. 1/- With less input views (the number of input views for the methods [17], [18] and our method are respectively 7, 25

and 5), our method outperforms [17] and achieves competitive results against [18]. 2/- Our method benefiting from the WTA fusion and the CNN based refinement is more robust and generates less artifacts for some light field scenes (c.f. Fig. 7). 3/- As [17], our method can predict disparity at each viewpoint, whereas the method [18] only predicts for the center view. 4/- Compared to EPI based learning frameworks, training our model is less demanding. Indeed, it is required in [18] to manually mark out and exclude all the reflection, refraction and textureless regions when preparing training data, which can be very time-consuming with a large dataset.

As learning-based methods may fail when characteristics in the input images differ from that in the training dataset, we have tested the robustness of our framework by adding Gaussian noise to the test light fields in comparison with two other learning-based methods [17], [18]. Fig. 8(a) shows the averaged MSE and Bad Pixel Ratio (threshold 0.07) over 8 synthetic LFs as a function of the standard deviation of the Gaussian noise. When increasing the standard variance of the noise, the performances of all the reference methods degrade, while the quality of depth maps estimated with the proposed framework remains more stable. To explain this difference in terms of robustness, we show in Fig. 8(b) 8(c), a clean and a polluted EPI. The added Gaussian noise destroys the geometric structures in the EPI that are used by methods like [17], [18] for depth estimation, while our framework exploits spatial information of each sub-aperture image, hence stays more robust against noise. Apart from the above experimental results on synthetic data, Fig. 10 also shows disparity maps estimated from light fields captured by plenoptic cameras, hence that are prone to noise and distortions. Fig. 10 shows that our framework can still estimate satisfying disparity maps.

A general flexibility and complexity comparison is summarized in Table III for all the compared methods. The proposed framework can adapt to light fields with wide baselines and, unlike methods using plane sweep volumes, does not require a discretized depth range at the input. Furthermore, some methods may be limited by their specific viewpoints selection pattern and cannot be used for estimated disparity maps for views located at the border of the light field. In contrast, the proposed approach uses a flexible stereo and anchor view selection pattern that allows us to estimate disparity maps for all light field views. In terms of computational cost, our framework takes less than 2 seconds to estimate one disparity map, that is much faster than traditional methods, but a bit slower than the two learning-based methods [17], [18]. Note that the implementations of the methods [6], [17] estimate disparity maps for all the views in a light field or for views on one row at one time, consequently we divided their costs by the number of estimated views. Since the codes of four traditional methods are not available for GPU, they are tested on an Intel i7 CPU with 16G RAM, whereas the three learning-based methods are tested on a NVIDIA Tesla P100 GPU with 16G memory.

Additionally, Table IV gives the contribution of each building block to the performance of the proposed framework. Thanks to our new dataset and our new finetuning strategy, *FN2-ft-stereo* significantly improves the accuracy of the estimated disparity maps compared with the original *FN2* model. By taking into account multiple anchor views at different viewpoints (the occluded regions for these views are unlikely all overlapping), the disparity fusion step is able to cope with errors in occluded regions. And the refinement aims at coping with disparity discontinuities that may be introduced by the fusion step. However, due to the fact that the occluded regions have a much smaller pixel number compared to the whole image, the fusion and refinement steps bring less quantitative improvement on the entire disparity map than the stereo estimation step.



(a) Evolution of the MSE and BP (threshold = 0.07) measures obtained with three learning-based estimation methods when increasing the standard deviation of Gaussian noise added to the input light fields $\sim\mathcal{N}(0,\,\sigma^2)$.

(b) EPI without Gaussian noise

(c) EPI with additive Gaussian noise $\sim\mathcal{N}(0,\,0.02^2)$

Fig. 8. (a) Impact of noise on the quality (in terms of MSE and BP) of estimated depth maps. (b) EPI from scene *stilllife* without Gaussian noise. (c) Same EPI but with Gaussian noise.

TABLE III
FLEXIBILITY COMPARISON

Property	[4]	[2]	[5]	[6]	[17]	[18]	Ours
Adaptability to wide baselines	×	×	√	√	×	×	√
Estimation for any view	×	×	√	√	√	×	√
Without disparity discretization	×	×	×	√	√	√	√
Computational cost (one view)	960s	>1h	127s	16s	0.04s	0.52s	1.93s

TABLE IV
CONTRIBUTION OF EACH BLOCK IN FRAMEWORK

Processing step	*FN2*	*FN2-ft-stereo*	Fusion	Refinement
MSE	10.94	2.27	2.00	1.91
BP	90.5	63.6	56.2	54.4

*MSE denotes 100*Mean Square Error, BP denotes Bad Pixel Ratios with threshold 0.01. The values for *FN2* and *FN2-ft-stereo* are averaged values of the estimations between 4 stereo views and the target view.

D. Results with Sparsely Sampled Synthetic Light Fields

Among the state-of-the-art approaches mentioned above, the methods in [2] and [4] derive the disparity estimate from EPI analysis, thus are hardly applicable for sparse light fields with large baselines. For sparsely sampled light fields, we compare our framework with the methods in [6], [5], [17] and [18], using the objective metrics: Mean Square Error (MSE) and Bad Pixel Ratios with larger thresholds 0.05, 0.1, 0.3. We consider the central 3×3 views of the light field, and the evaluation is performed for the estimated disparity of the central view. The step length in the method in [5] is set to 0.05, corresponding to the minimal Bad pixel ratio threshold. Both the input view number in [17] and the stream length in [18] are set to be 3. The rest of the setup is identical to the experiments for dense light fields. The lower parts of Table I and II show that our framework yields better Bad Pixel Ratios with large margins when compared with other methods. In terms of MSE, our method ranks the second, slightly lagging behind [5]. The first two columns in Fig. 9 show disparity maps estimated with different methods. Compared with the first two non deep learning-based methods, our algorithm functions well in both contours and homogeneous zones. Although our method gives smoother boundaries than those with [5], it works well in slanted zones where the method in [5] tends to fail. For the two deep learning-based methods [17], [18], even though we have re-trained the corresponding models with our sparse light field data, the results contain severe deformations and artifacts, since epipolar line continuity is no longer guaranteed with sparse light fields.

E. Results with Real Light Fields

To assess the disparity estimation performance of our network with real light fields, we consider both dense light fields (7×7 central views are considered) captured with the Lytro Illum plenoptic camera [41] [42] and sparse light fields (3×3 central views are considered) captured with camera arrays [43]. Because of the lack of ground truth disparity values, the prior disparity range required by these methods is set using our estimation results, with a margin of 10%: $[d_{min} - 0.05(d_{max} - d_{min}), d_{max} + 0.05(d_{max} - d_{min})]$. Fig. 10 shows the estimated disparity maps for dense light fields using different methods. Two sparse light fields are tested in Fig. 9. Compared with other methods, our framework gives more accurate estimates on object boundaries, especially for the scenes which have more texture details. Among all results, our disparity maps have less artifacts in spite of the fact that relatively less views are used for the estimation. Note that our network has been trained with only noise-free synthetic light fields. An additional fine-tuning with noisy images could still improve the estimation accuracy of our network for real light fields.

Overall, regardless their limitations in terms of view number and view selection, the methods in [4], [6] show inaccuracies at the boundaries, while methods in [2], [5] can give relatively more accurate estimates at the boundaries. The method in [2] often contains artifacts in homogeneous regions, while the method in [5] fails on slanted surfaces. Our experiments with noisy light fields have shown that the performance of the learning-based methods [17] and [18] depends on the quality of EPI. On the contrary, our method can estimate accurate depth maps both at the boundaries of objects, and in homogeneous and slanted regions with an acceptable computational cost. It has also been show to be more robust to noise than the other two learning-based methods.

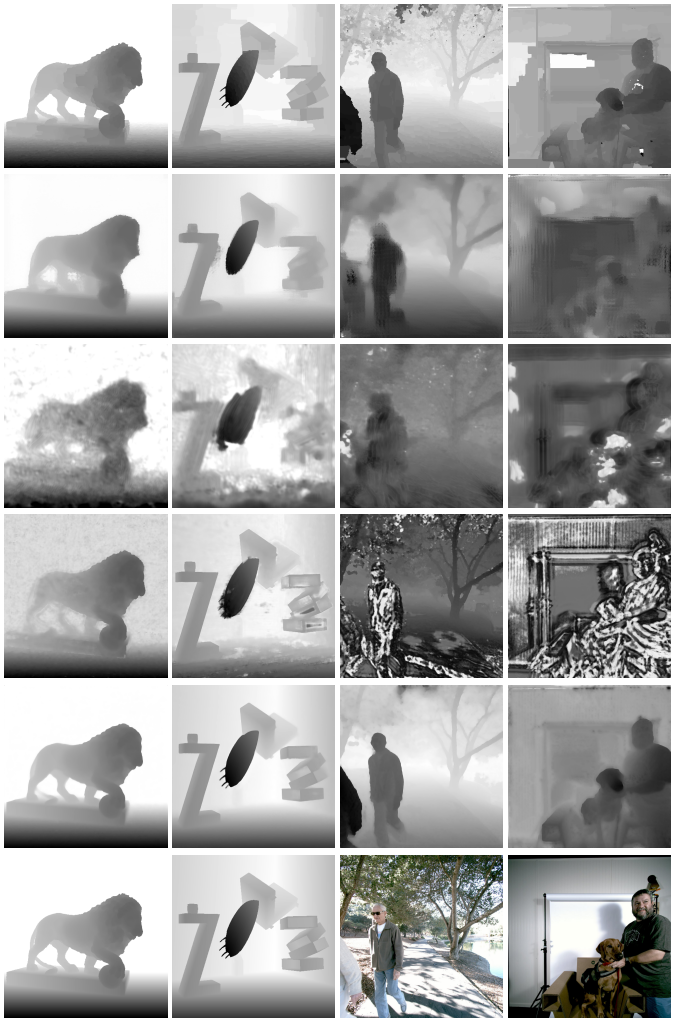


Fig. 9. Visual comparison for estimated disparity maps for sparse light fields. The first two columns are obtained with synthetic data, while the last two columns are obtained for real-world data. From top to the bottom: *Huang* [5], *Jiang et al.* [6], *Heber et al.* [17], *Shin et al.* [18], and ours. The final row is the ground truth disparity (when available) or the color image of the central viewpoint.

VIII. Conclusion

In this paper, we have proposed a learning based approach to estimate disparity maps between all light field views. The algorithm takes a variable subset of input views and estimates accurate disparity maps for both densely and sparsely sampled LF data.

The proposed framework starts with a stereo estimation step which takes a flexible number of light field views to give first disparity maps estimates. A fusion step then aggregates these disparity maps into a single one by conducting pixel-wise selection based on the warping error. A multiscale residual

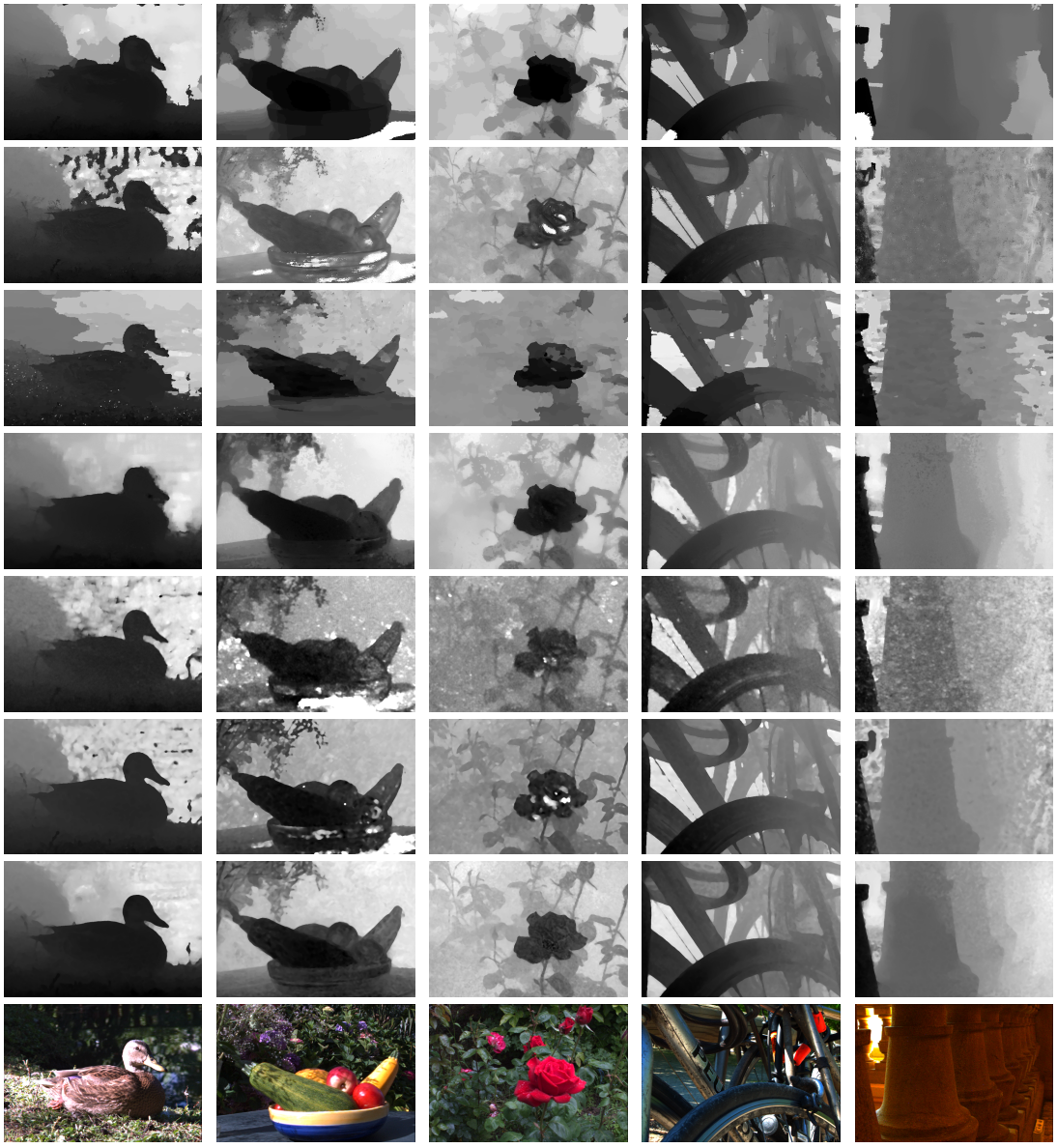


Fig. 10. Qualitative comparisons for estimated disparity maps. The first three light fields are from the INRIA Dataset [41], and the last two light fields are from the EPFL dataset [42]. From top to bottom, figures show the disparity maps estimated with methods in *Jeon et al.* [4], *Zhang et al.* [2], *Huang* [5], *Jiang et al.* [6], *Heber et al.* [17], *Shin et al.* [18] and our proposed method. The final row shows the central views of the light fields.

refinement step is then used to eliminate noise and improve spatial coherence. In order to train the model so that it can apply to both sparsely and densely sampled light fields, we have also created two synthetic light fields datasets with different disparity ranges. To our knowledge, this is the first publicly available dataset for sparsely sampled synthetic light fields given together with ground truth disparity maps for all the views.

The effectiveness of our algorithm has been demonstrated with both synthetic and real light fields datasets, in comparison with several state-of-the-art reference methods. The proposed algorithm outperforms state-of-the-art algorithms despite of the use of less input views. It is robust in both homogeneous areas and along the contours, as well as in slanted zones. Experimental results with real light fields show that our algorithm estimates consistent objects boundaries, and preserves details in the scene, although the network has been only trained using synthetic data.

REFERENCES

[1] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 606–619, Aug. 2013.

[2] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *J. of Computer Vision and Image Understanding*, vol. 145, pp. 148–159, Apr. 2016.

[3] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *Int. J. of computer vision*, vol. 1, no. 1, pp. 7–55, Mar. 1987.

[4] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. So Kweon, "Accurate depth map estimation from a lenslet light field camera," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1547–1555.

[5] C.-T. Huang, "Empirical bayesian light-field stereo matching by robust pseudo random field modeling," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 1–1, Feb. 2018.
[6] X. Jiang, M. Le Pendu, and C. Guillemot, "Depth estimation with occlusion handling from a sparse set of light field views," in *IEEE Int. Conf. on Image Processing (ICIP)*, 2018, pp. 634–638.
[7] J. Navarro and A. Buades, "Robust and dense depth estimation for light field images," *IEEE Trans. on Image Processing*, vol. 26, no. 4, pp. 1873–1886, Apr. 2017.
[8] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2015, pp. 2758–2766.
[9] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1647 – 1655.
[10] E. Ilg, T. Saikia, M. Keuper, and T. Brox, "Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation," in *Eur. Conf. on Computer Vision (ECCV)*, 2018, pp. 626–643.
[11] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2720 – 2729.
[12] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8934–8943.
[13] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. on Graphics*, vol. 35, no. 6, pp. 193:1–193:10, 2016.
[14] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow." in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 4473–4481.
[15] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan, "LFNet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution," *IEEE Trans. on Image Processing*, vol. 27, no. 9, pp. 4274–4286, May 2018.
[16] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate superresolution," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5835–5843.
[17] S. Heber, W. Yu, and T. Pock, "Neural EPI-volume networks for shape from light field," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 2271–2279.
[18] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, "EPINET: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4748–4757.
[19] X. Jiang, J. Shi, and C. Guillemot, "A learning based depth estimation framework for 4D densely and sparsely sampled light fields," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2257–2261.
[20] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Asian Conf. on Computer Vision (ACCV)*, 2016, pp. 19–34.
[21] O. Johannsen, K. Honauer, B. Goldluecke, A. Alperovich, F. Battisti, Y. Bok, M. Brizzi, M. Carli, G. Choe, M. Diebold *et al.*, "A taxonomy and evaluation of dense light field depth estimation algorithms." in *IEEE Conf. on Computer Vision and Pattern Recognition workshop (CVPR workshop)*, 2017, pp. 1795–1812.
[22] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, Dec. 2002.
[23] O. Veksler, "Stereo matching by compact windows via minimum ratio cycle," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2001, pp. 540–547.
[24] C. L. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 7, pp. 675–684, Jul. 2000.
[25] T. Taniai, Y. Matsushita, and T. Naemura, "Graph cut based continuous stereo matching using locally shared labels," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1613–1620.
[26] D. Maurer, Y. C. Ju, M. Breuß, and A. Bruhn, "Combining shape from shading and stereo: A joint variational method for estimating depth, illumination and albedo," *Int. J. of Computer Vision*, vol. 126, no. 12, pp. 1342–1366, Dec. 2018.
[27] L. Zhang and S. M. Seitz, "Estimating optimal parameters for MRF stereo from a single image pair," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 331–342, Feb. 2007.
[28] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–8.
[29] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1592–1599.
[30] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang, "A deep visual correspondence embedding model for stereo matching costs," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2015, pp. 972–980.
[31] A. Kendall, H. Martirosyan, S. Dasgupta, and P. Henry, "End-to-end learning of geometry and context for deep stereo regression," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 66–75.
[32] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4040–4048.
[33] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *IEEE Int. Conf. on Computer Vision workshops (ICCV workshop)*, 2017, pp. 878 – 886.
[34] "MIT synthetic light field archive," http://web.media.mit.edu/~gordonw/SyntheticLightFields/index.php.
[35] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," in *Conf. on Vision, Modeling & Visualization (VMV)*, 2013, pp. 225–226.
[36] "Blender website," https://www.blender.org/.
[37] "Chocofur website," http://www.chocofur.com/.
[38] "Sketchfab website," https://sketchfab.com/.
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computing Research Repository*, vol. abs/1412.6980, 2014.
[40] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning," in *USENIX Symp. on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.
[41] X. Jiang, M. Le Pendu, R. A. Farrugia, and C. Guillemot, "Light field compression with homography-based low-rank approximation," *J. of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1132–1145, Oct. 2017.
[42] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Int. Conf. on Quality of Multimedia Experience (QoMEX)*, no. EPFL-CONF-218363, 2016.
[43] E. Penner and L. Zhang, "Soft 3D reconstruction for view synthesis," *ACM Trans. on Graphics*, vol. 36, no. 6, pp. 235:1–235:11, 2017.

Jinglei Shi, received Bachelor's degree in Electronic Information Engineering from UESTC (University of Electronic Science and Technology of China), China, in 2015. Then he received Engineer's degree and Master's degree in Image Processing from IMT (Institut Mines-Télécom) Atlantique, France, in 2017. He is currently a PhD student at INRIA (Institut National de Recherche en Informatique et en Automatique) in France. His research interests concern light field imaging and learning-based depth estimation and view synthesis.

Xiaoran Jiang, received the Engineering degree in Telecommunications and the Ph.D. degree in Neural Networks from IMT (Institut Mines-Télécom) Atlantique, France, in 2010 and 2014, respectively. He is currently a Research Fellow at INRIA (Institut National de Recherche en Informatique et en Automatique) in France. His current research interests include signal and image processing, computer vision, and in particular machine learning methods for light field compression, view synthesis and depth estimation.

Christine Guillemot, IEEE fellow, is Director of Research at INRIA, head of a research team dealing with image and video modeling, processing, coding and communication. She holds a Ph.D. degree from ENST (Ecole Nationale Superieure des Telecommunications) Paris, and an Habilitation for Research Direction from the University of Rennes. From 1985 to Oct. 1997, she has been with FRANCE TELECOM, where she has been involved in various projects in the area of image and video coding for TV, HDTV and multimedia. From Jan. 1990 to mid 1991, she has worked at Bellcore, NJ, USA, as a visiting scientist. Her research interests are signal and image processing, and in particular 2D and 3D image and video processing for various problems (compression, super-resolution, inpainting, classification).

She has served as Associate Editor for IEEE Trans. on Image Processing (from 2000 to 2003, and from 2014-2016), for IEEE Trans. on Circuits and Systems for Video Technology (from 2004 to 2006), and for IEEE Trans. on Signal Processing (2007-2009). She has served as senior member of the editorial board of the IEEE journal on selected topics in signal processing (2013-2015) and is currently senior area editor of IEEE Trans. on Image Processing.